Dataset Selection

Knowing I would have limited time to complete this project (and being very new to the world of machine learning), I wanted to select a dataset which would allow me to learn the fundamentals of ML quickly and effectively while also having an interesting real-world application. After researching available datasets, it was clear to me that the Iris dataset is widely considered to be the equivalent of the “Hello World” program in the ML world. Being a small dataset (with 5 columns and 150 rows), it was clear no input sanitation would need to be performed on the dataset. I would be able to get right into training the model with a variety of different supervised ML models without worrying if the dataset has been optimized. The objective of the dataset is also very clear – simply determining the type of flower based on its attributes. The objective of the analysis is interesting and slightly complex as it is not simply a Boolean result (either yes or no), but rather segmented into three different flower classes which need to be identified.

Model Selection

A variety of models were compared against each other after determining their accuracy using the cross-validation technique (the process of which is covered in detail within the readme). The models tested included:

Model	Average Evaluation Score (%)
Logistic Regression	95.6%
Decision Tree Classifier	95.1%
K-Nearest Neighbors Classifier	95.1%
Gaussian Naïve Bayes	95.1%
Linear Discriminant	97.6%
Support Vector Machines	98.3%

The Support Vector Classification model achieved the highest average evaluation score of 98.3%, making it the most appropriate model to use for this dataset. It also achieved a prediction accuracy of 96.7% on the test data, confirming it to be a well performing model (also avoiding the perils of overfitting and underfitting).